Standardised Stops: The Bias Factor

Ahana Yelagar, Amelia Burnell, Grace Bero, Bridget Brombach

Introduction

- We were hired by - the Durham County Police Department.

- Our goal is to see if there is any bias in issuing citations in the years 2014-2015.

- The dataset we used includes information about various aspects of a police encounter (e.g: the reason for the stop, demographic data, outcomes, etc.)

Data Cleaning Steps

01 Filtered out the data for the years 2014 & 2015 only

02 Created categories based on time of day

03 Added a column to indicate if the event happened in the first three weeks of the month.

04 Added a column for which day of the week the event happened

05 Cleaned the “age” column to exclude NA entries

Complete Separation Testing

01	**Made Table Between All Categorical Xs, and Binary Y**

No empty (zero) cells

02	**Made a Scatterplot between numeric X and Binary Y**

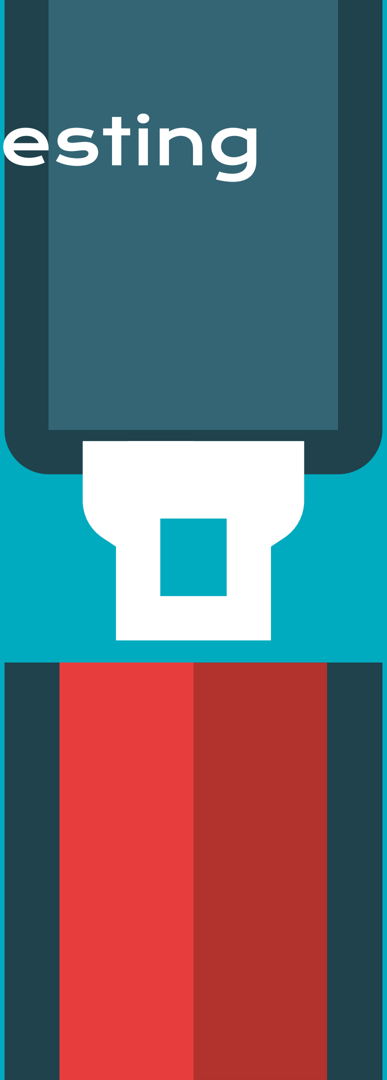
Looks evenly distributed/ no obvious separation

03	**Checking Standard Error of Betas**

None of the Beta Standard Errors > 0.5

04	**Conclusion**

3 of 3 tests indicate no complete separation. We will be using maximum likelihood analysis



Our Model

$$Y \sim Bernoulli(\pi)$$

$$log\left(\frac{\pi_i}{1\text{-}\pi_i}\right) = \beta_0 + \beta_{\text{age\_numeric}} * (age\_\text{numeric}) + \beta_{asian/pacific\ islander} * (asian/pacific\ islander)$$

$$+ \beta_{\text{black}} * (\text{black}) + \ldots + \beta_{\text{stop light/ sign violation}} * (Stop\ Light/Sign\ Violation)$$

Why are we using a logit link function? Why not just use ORL?

1. Y is binary - you can either get a ticket, or not
2. Least squares will not work
3. A linear model allows for probabilities <0 and >1.

Analysis of Maximum Likelihood Estimates				
Parameter		DF	Estimate	Standard Error
Intercept		1	-0.5490	0.1544
age_numeric		1	-0.0104	0.000762
subject_race	asian/pacific islander	1	0.1192	0.0787
subject_race	black	1	-0.00611	0.0232
subject_race	hispanic	1	0.7542	0.0367
subject_race	other	1	0.3179	0.1270
subject_race	unknown	1	-0.0714	0.1424
time_cat	Afternoon	1	0.3728	0.1481
time_cat	Early Morn	1	0.5280	0.1514
time_cat	Evening	1	0.4610	0.1488
time_cat	Late Night	1	0.0264	0.1496
time_cat	Morning	1	0.5647	0.1481
time_cat	Night	1	0.1489	0.1481
weekday	1	1	-0.0216	0.0465
weekday	2	1	-0.0463	0.0429
weekday	3	1	0.0285	0.0391
weekday	4	1	0.1165	0.0372
weekday	5	1	0.0730	0.0373
weekday	6	1	0.0805	0.0381
time_of_month	first3	1	-0.0397	0.0210

Analysis of Maximum Likelihood Estimates				
Parameter		DF	Estimate	Standard Error
subject_sex	female	1	0.0189	0.0205
reason_for_stop	Checkpoint	1	1.1999	0.0562
reason_for_stop	Driving While Impaired	1	-1.7033	0.2865
reason_for_stop	Investigation	1	-0.5295	0.0522
reason_for_stop	Other Motor Vehicle Violation	1	-0.2575	0.0717
reason_for_stop	Safe Movement Violation	1	-0.8535	0.0432
reason_for_stop	Seat Belt Violation	1	1.2547	0.0571
reason_for_stop	Speed Limit Violation	1	0.8634	0.0267
reason_for_stop	Stop Light/Sign Violation	1	0.2722	0.0396
reason_for_stop	Vehicle Equipment Violation	1	-0.8794	0.0405

Testing the model as a whole

$$H_0: \beta_0 = \beta_{\text{age}} = \beta_{asian/pacific\ islander} = \beta_{\text{black}} = \ldots = \beta_{\text{stop light/ sign violation}}$$

$$H_A: \text{at least one } \beta \text{ is not equal to } 0$$

$$l_r = 6801.4883$$

$$p - value =< 0.0001$$

$$\text{null distribution: } \chi^2(29)$$

$$\alpha = 0.01$$

Our p-value < alpha. We reject the null hypothesis.

We do have evidence that at least one coefficient is not equal to 0.

Age bias

$$H_0: \beta_{\text{age\_numeric}} = 0$$
$$H_A: \beta_{\text{age\_numeric}} \neq 0$$

$$\omega = 186.8371$$
$$p - value =< 0.0001$$
$$\text{null distribution: } \chi^2(1)$$
$$\alpha = 0.01$$

Our p-value < alpha. We reject the null hypothesis.
We do have evidence that age is a significant predictor of a citation being issued.

Race Bias

$$H_0: \beta_{Asian/Pacific\ Islander} = \beta_{Black} = \beta_{Hispanic} = \beta_{Other} = \beta_{Unknown} = 0$$

$$H_a: At\ least\ one\ \beta \neq 0$$

$$Test\ Statistic: \omega = 539.5991$$

$$Null\ Distibution: \chi^2(5)$$

$$p - value: < .0001$$

$$\alpha = 0.01$$

Our p-value < alpha. We **reject** the null hypothesis.
We **do** have evidence that **race is a significant predictor of a citation being issued.**

Significance of Variables: Summary

Variable	Categories	Significance
age_numeric	Quantitative	Significant
subject_sex	Female, male	Significant
subject_race	asian/pacific islander, black, hispanic, other, unknown, white	Significant
time_cat	Afternoon, early morn, evening, late night, morning, night, other	Significant
weekday	1, 2, 3, 4, 5, 6, 7	Significant
time_of_month	First, Last Week of Month	Not Significant
reason_for_stop	Checkpoint, Driving while impaired, investigation, other motor vehicle violation, safe movement violation, seat belt violation, speed limit violation, stop light/sign violation, vehicle equipment violation, vehicle regulatory violation	Significant

Personas - black female

Consider a black female, named Jane Doe, who is 20 years old and was stopped on a saturday in the last week of the month, at an unknown time. What is the estimated probability that she gets a ticket?

Linear Predictor

$$\eta = -0.5490 - 0.0104(20) - 0.00611(1) + 0.0189(1)$$
$$= -0.7992$$

Estimated Probability

$$P = \frac{e^{-0.7992}}{1+e^{-0.7992}} = 0.3102$$

Conclusion: The estimated probability of a **20-year-old black female** getting a ticket, is approximately **0.3102, or 31.02%.**

JANE DOE

XX.XXXX.XXX

XX DRIVER LICENSE XX

$$\beta_{\text{black}} = 1$$
$$\beta_{\text{age\_numeric}} = 20$$
$$\beta_{\text{female}} = 1$$
all other β are equal to 0

Personas - white male

Consider a white male, named John Smith, who is also 20 years old and was stopped on a Saturday in the last week of the month, at an unknown time. What is the estimated probability that he gets a ticket?

Linear Predictor

$$\eta = -\,0.5490 - 0.0104(20)$$
$$= -0.757$$

Estimated Probability

$$P = \frac{e^{-0.757}}{1+e^{-0.757}} = 0.3129$$

Conclusion: The estimated probability of a **20-year-old white male** getting a ticket, is approximately **0.3129, or 31.29%.**

JOHN SMITH

XX.XXXX.XXX

XX DRIVER LICENSE XX

$$\beta_{white} = 0$$
$$\beta_{age\_numeric} = 20$$
$$\beta_{male} = 0$$
all other β are equal to 0

Calculating the odds ratio of a black female receiving a ticket over a white male -

$$\text{odds ratio} = \frac{\text{odds for a black female}}{\text{odds for white male}} = \frac{e^{\beta o + \beta black + \beta age(Age) + \beta female}}{e^{\beta o + \beta age(Age)}}$$

$$= \frac{e^{-0.5547-0.00628-0.0104(20)+0.0188}}{e^{-0.5547-0.0104(20)}} = \frac{0.4724}{0.4664} = 1.012$$

Holding all other explanatory variables constant, being a black female is associated with an increase by 1.2% in odds of receiving a ticket over a white male.

Jane Doe has slightly higher odds of receiving a ticket over John Smith

Odds Ratio - Subject Age

How would the **odds** of receiving a citation change for a **decrease in age of 10 years?**

$$e^{-10*\beta_{\text{age\_numeric}}} = e^{-10*-0.0104} = 1.1096$$

A person who is 20 has **10% increase in odds** of being issued a citation compared to someone of the same profile who is 30.

Older is better if you don't want to get a citation!

Confidence intervals

- We are 95% confident that the **odds** of receiving a citation change **by a factor between** 0.9496 and 1.039 for a **person identifying as female over a person who identifies as male**, holding all other variables constant. In other words, we are 95% sure that the odds **don't change that much**.
- We are 95% confident that the **odds** of receiving a citation increase **by a factor between** 1.978 and 2.285 for a **person identifying as hispanic over a person identifying as white**, holding all other variables constant. In other words, we are 95% sure that the odds almost **double.**

Conclusions

01 The only variable that is not significant is the time of month

02 Based on our two personas, there looks to be only a slight bias in odds of receiving a citation based on gender and identifying as black vs white

03 The older the subject, the lower the odds of them getting a ticket

04 There does appear to be racial bias,based on our evaluation of the data

05 There is significant bias for a white identifying individual compared to a hispanic identifying individual, odds of receiving citation approx. doubled

Thank you!